

Parsimonious Skew Mixture Models for Model-Based Clustering and Classification

Irene Vrbik and Paul D. McNicholas*

Department of Mathematics & Statistics, University of Guelph.

Abstract

In recent work, robust mixture modelling approaches using skewed distributions have been explored to accommodate asymmetric data. We introduce parsimony by developing skew- t and skew-normal analogues of the popular GPCM family that employ an eigenvalue decomposition of a positive-semidefinite matrix. The methods developed in this paper are compared to existing models in both an unsupervised and semi-supervised classification framework. Parameter estimation is carried out using the expectation-maximization algorithm and models are selected using the Bayesian information criterion. The efficacy of these extensions is illustrated on simulated and benchmark clustering data sets.

1 Introduction

The objective of cluster analysis is to organize data into groups wherein the similarity within groups and the dissimilarity between groups are maximized. A ‘model-based’ approach is one that uses mixture models for clustering. The use of finite mixture models has become increasingly common for clustering and classification, particularly with the use of Gaussian components. The Gaussian model-based clustering likelihood is

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{j=1}^n \sum_{i=1}^g \pi_i \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (1)$$

where $\pi_i > 0$, such that $\sum_{i=1}^g \pi_i = 1$, are mixing proportions and $\phi(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the density of a multivariate Gaussian random variable with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Model-based classification is a semi-supervised version of model-based clustering (cf. Section 6).

Gaussian mixture models have been used for a wide variety of clustering applications, including work by McLachlan and Basford (1988), Bouveyron et al. (2007), McNicholas and Murphy (2008, 2010a,b), and Baek and McLachlan (2010), amongst others. In efforts to accommodate data that exhibit some departure from normality, robust extensions are garnering increased attention. For instance, mixtures of multivariate t -distributions (McLachlan and Peel, 1998; Peel and McLachlan, 2000) have proven effective for dealing with components containing outliers. They have been the basis of a variety of robust clustering techniques that use mixtures of multivariate t -distributions, including work by McLachlan et al. (2007), Andrews and McNicholas (2011a,b), Baek and McLachlan (2011), Steane et al. (2012), and McNicholas and Subedi (2012), amongst others.

Capturing components that are asymmetric can be tackled using skew-normal distributions (cf. Lin et al., 2007) or other non-elliptically contoured distributions (e.g., Karlis and Santourian, 2009). One example of a non-elliptical distribution is the skew-normal independent distribution, as considered for finite mixture

*E-mail: paul.mcnicholas@uoguelph.ca. Tel: +1-519-824-4120, ext. 53136.

modeling in Cabral et al. (2012). Another interesting alternative is the skew Student- t -normal distribution that has recently been used to model skewed heavy-tailed data (Ho, Pyne, and Lin, 2012). Although there are many non-symmetric options available, we focus on mixtures of skew-normal distributions and mixtures of skew- t distributions herein.

Recently, mixtures of multivariate skew- t distributions have been receiving some attention in the literature. Of course, the skew-normal, t , and Gaussian distributions are all special cases of the skew- t distribution. This property can be important in clustering applications because we often do not know the most appropriate underlying distribution. Alongside the skewness parameter that accommodates asymmetric data, the degrees of freedom parameter allows for heavy tails, giving less weight to outlying observations in parameter estimation. Compared to work on parsimonious Gaussian and t -mixtures, the literature contains relatively little on parsimonious mixtures of multivariate skew-normal and skew- t distributions. The purpose of this paper is to go some way towards addressing this deficiency.

The remainder of the paper is organized as follows. In Section 2, we introduce skew- t and skew-normal mixture models and briefly discuss the calculation of parameter estimates. Section 3 presents the construction of parsimonious families of models that are analogues of popular Gaussian approaches. The proposed methods are compared with Gaussian and multivariate t analogues using simulation studies (Section 4) and four benchmark clustering data sets (Section 5). These models are extended further to semi-supervised classification in Section 6, which is followed by concluding remarks (Section 7).

2 Mixtures of Skew- t and Skew-Normal Distributions

2.1 A Mixture of Skew- t Distributions

Although the computational tractability of Gaussian mixture models has contributed to their widespread popularity within the literature, their application is not always appropriate. For instance, Kotz and Nadarajah (2004) argue that the multivariate- t distribution provides a more realistic model for real-world data and it has been noted (e.g., Lin et al., 2007) that Gaussian mixture models have a tendency to over fit skewed data. Therefore, it is natural to consider a single distribution, namely the multivariate skew- t distribution, that conflates the robust properties of the t -distribution with a skewness parameter to account for asymmetry. We outline a model-based approach using parsimonious mixtures of multivariate skew- t distributions; we also consider a mixture of skew-normal distributions, which is a limiting case.

There are a number of ‘skew- t ’ distributions within the literature. The version that we adopt, as defined by Pyne et al. (2009), uses a particular stochastic representation of the multivariate skew- t distribution given by Sahu et al. (2003). This distribution also corresponds to the so-called ‘restricted’ multivariate skew- t distribution, as defined in Lee and McLachlan (2012). Adopting this characterization, a random vector \mathbf{Y} is said to follow a p -variate skew- t distribution with location vector $\boldsymbol{\xi}$, scale matrix $\boldsymbol{\Omega}$, skewness vector $\boldsymbol{\lambda}$, and ν degrees of freedom if it has the representation

$$\mathbf{Y} = \boldsymbol{\lambda}|U| + \mathbf{X}, \quad (2)$$

where

$$\begin{pmatrix} \mathbf{X} \\ U \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \boldsymbol{\xi} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \frac{1}{w} \right),$$

and $W \sim \Gamma(\nu/2, \nu/2)$.

We consider a g -component mixture of p -dimensional skew- t distributions with density given by

$$f(\mathbf{y}_j | \boldsymbol{\Psi}) = \sum_{i=1}^g \pi_i \mathcal{K}(\mathbf{y}_j | \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \nu_i), \quad (3)$$

where $\kappa(\mathbf{y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \nu_i)$ is the density of a multivariate skew- t distribution with location vector $\boldsymbol{\xi}_i$, scale matrix $\boldsymbol{\Omega}_i$, skewness parameter $\boldsymbol{\lambda}_i$, and ν_i degrees of freedom, and Ψ contains all model parameters, i.e., Ψ contains the parameters $\{\pi_i, \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \nu_i : i = 1, \dots, g\}$.

2.2 A Mixture of Skew-Normal Distributions

Now consider the skew-normal distribution, which is a limiting case of the skew- t distribution. This version of the skew-normal distribution is the ‘restricted’ version of the skew-normal distribution defined by Sahu et al. (2003) and proposed for the analysis of flow cytometric data by Pyne et al. (2009). Resembling the above characterization (Section 2.1), a random vector \mathbf{Y} is said to follow a p -variate skew-normal distribution with location vector $\boldsymbol{\xi}$, scale matrix $\boldsymbol{\Omega}$, and skewness vector $\boldsymbol{\lambda}$ if it has the representation

$$\mathbf{Y} = \boldsymbol{\lambda}|U| + \mathbf{X}, \quad (4)$$

where

$$\begin{pmatrix} \mathbf{X} \\ U \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\xi} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}\right),$$

We consider a g -component mixture of p -dimensional skew-normal distributions. The density is given by

$$f(\mathbf{y}_j \mid \boldsymbol{\Theta}) = \sum_{i=1}^g \pi_i \varphi(\mathbf{y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i), \quad (5)$$

where $\varphi(\mathbf{y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \nu_i)$ is the density of a multivariate skew-normal distribution with location vector $\boldsymbol{\xi}_i$, scale matrix $\boldsymbol{\Omega}_i$, and skewness $\boldsymbol{\lambda}_i$, and $\boldsymbol{\Theta}$ contains all model parameters, i.e., $\boldsymbol{\Theta}$ contains $\{\pi_i, \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i : i = 1, \dots, g\}$.

2.3 Parameter Estimation

Parameter estimation for mixtures of multivariate skew-normal distributions is carried out via an expectation-maximization (EM) algorithm (Dempster et al., 1977) as outlined by Pyne et al. (2009). However, maximum likelihood estimation for mixtures of multivariate skew- t distributions is more involved and has been tackled using a number of different variations of the EM algorithm. Vrbik and McNicholas (2012) derive closed form solutions for the skew- t random variable defined in (2), allowing for a traditional EM algorithm to be used. This aforementioned procedure is employed for our proposed study but multiple techniques are available. For example, a Bayesian approach is explored in Frühwirth-Schnatter and Pyne (2010), wherein an efficient Markov chain Monte Carlo (MCMC) scheme is proposed for finite mixtures of multivariate skew-normal and skew- t distributions.

A nice overview of the differing characterizations of the skew- t distribution and their applications has recently been given by Lee and McLachlan (2012). Apart from providing an up-to-date account of the recent work with the skew- t distribution, they also develop some new results on the restricted and unrestricted multivariate skew- t distribution (rMST and MST, respectively). The E-step for mixtures of MST can be expressed in closed form apart from one term and, when using a one-step-late M-step, this term can be taken to be zero or be expressed in closed form. Lee and McLachlan (2012) offer an alternative method to calculate the intractable E-step for the rMST. This is accomplished by manipulating the expectations to be expressed in terms of the moments of the truncated non-central multivariate t -distribution (cf. Ho, Lin, Chen, and Wang, 2012).

2.4 Comments

The additional skewness parameter gives these models the advantage of being able to adjust for asymmetric data while the degrees of freedom in the skew- t case acts as a robustness tuning parameter to accommodate

heavier tails. The contour maps in Figure 1 illustrate various shapes attained by the bivariate skew-normal distribution, with teardrop-like as well as spherical and elliptically shaped clusters. The pictures are constructed using a multivariate skew-normal distribution with $\mathbf{\Omega} = \mathbf{I}$ to isolate the effect of skewness on otherwise elliptically shaped clusters. Note that a multivariate skew- t distribution with large degrees of freedom would appear similar, but we would see more in the tails with lower values of degrees of freedom.

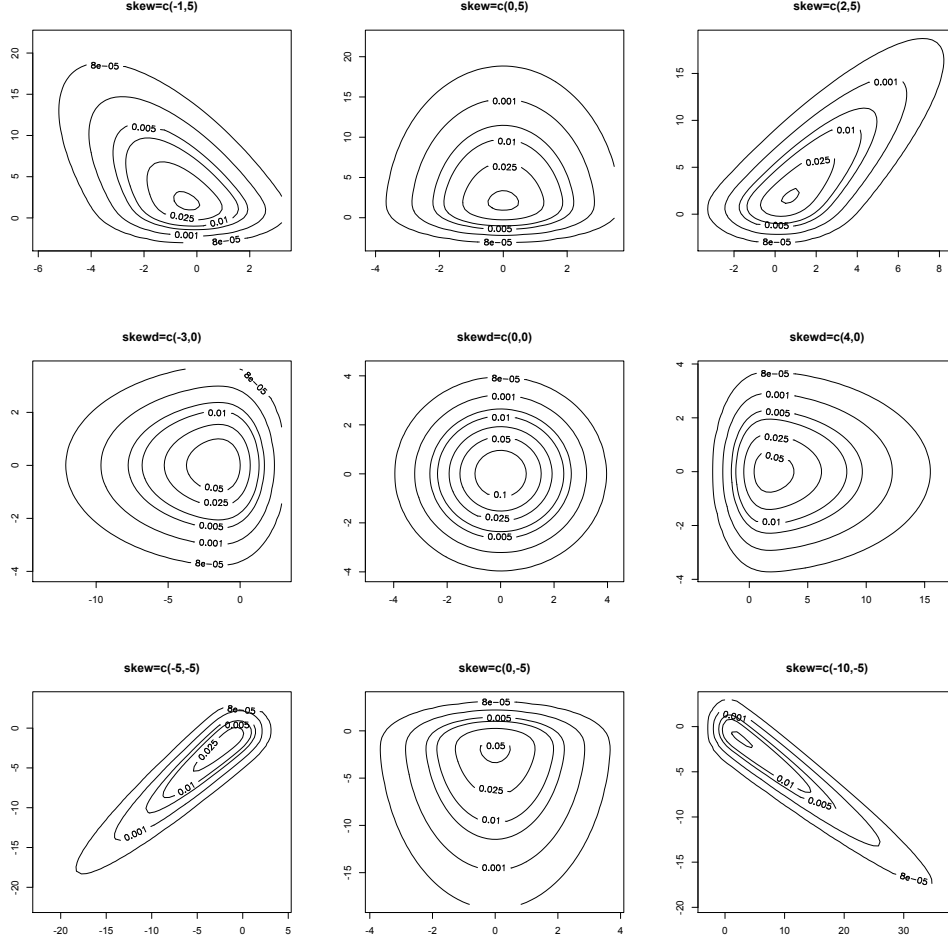


Figure 1: Density contours from bivariate skew-normal distributions with $\boldsymbol{\xi} = \mathbf{0}$, $\mathbf{\Omega} = \mathbf{I}$, and varying values for the skewness parameter ($\boldsymbol{\lambda}$).

3 Methodology

3.1 Two Skewed Families

Despite the fact that there has been support for model-based approaches for quite some time now, only in the last two or more decades has this approach become popular. Owing greatly to the advances in computing, model-based approaches are quickly becoming an attractive tool for clustering and classification.

For Gaussian model-based clustering, model fitting becomes an issue due to the number of parameters in the component covariance matrices: there are $Gp(p+1)/2$ in all, cf. (1). This same problem manifests for mixtures of multivariate t , skew-normal, and skew t -distributions.

Celeux and Govaert (1995) introduced parsimony into Gaussian mixtures by imposing constraints on eigen-decomposed component covariance matrices. This parameterization, originally considered in Banfield and Raftery (1993), is an eigenvalue decomposition of the component covariance matrices given by $\Sigma_i = \lambda_i \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i'$, where \mathbf{D}_i is the orthogonal matrix of eigenvectors of Σ_i , \mathbf{A}_i is the diagonal matrix of entries proportional to eigenvalues with $|\mathbf{A}_i| = 1$, and λ_i is the associated constant of proportionality. Celeux and Govaert (1995) developed a family of fourteen Gaussian parsimonious clustering models (GPCMs) by applying different constraints to this eigen-decomposed covariance structure (Table 1). A subset of these models make up the of family ten models available in the `mclust` package (Fraley et al., 2012) for R (R Development Core Team, 2012). In addition to clustering applications, Gaussian mixture models with this covariance decomposition have been used for discriminant analysis (Bensmail and Celeux, 1996) and classification (Dean et al., 2006).

Table 1: The nomenclature used in the MCLUST family and borrowed for the models developed herein, along with the decomposition constraints and number of free parameters in the decomposed matrix. Note: ‘E’ denotes equal across groups, ‘V’ denotes varying across groups, ‘I’ denotes the identity matrix.

Model	$\lambda_i = \lambda$	$\mathbf{A}_i = \mathbf{A}$	$\mathbf{D}_i = \mathbf{D}$	Free Covariance/Scale Parameters
EII	E	I	I	1
VII	V	I	I	G
EEI	E	E	I	p
VEI	V	E	I	$G + (p - 1)$
EVI	E	V	I	$Gp - (G - 1)$
VVI	V	V	I	Gp
EEE	E	E	E	$[p(p+1)/2]$
VEE*	V	E	E	$[p(p+1)/2] + (G - 1)$
EVE* ⁺	E	V	E	$[p(p+1)/2] - (G - 1)(p - 1)$
VVE* ⁺	V	V	E	$[p(p+1)/2] + (G - 1)p$
EEV	E	E	V	$G[p(p+1)/2] - (G - 1)(p)$
VEV	V	E	V	$G[p(p+1)/2] - (G - 1)(p - 1)$
EVV*	E	V	V	$G[p(p+1)/2] - (G - 1)$
VVV	V	V	V	$G[p(p+1)/2]$

*These models are not considered in MCLUST.

+These models are not considered in *t*EIGEN.

Andrews and McNicholas (2012a) introduced the *t*EIGEN family, which comprises *t*-analogues of the MCLUST models, along with two other GPCM models, with the added restriction of constraining or not constraining the component degrees of freedom ν_i to be equal across groups. The *t*EIGEN family is supported by the `tEigen` package (Andrews and McNicholas, 2012b) for R. The nomenclature for the *t*EIGEN models consist of four letters: ‘C’ denotes that a constraint is imposed, ‘U’ denotes that a constraint is not imposed, and ‘I’ denotes that the matrix in question is taken to be the identity matrix of suitable dimension. For each member of the *t*EIGEN family, relevant letters indicate the constraints on λ_i , \mathbf{D}_i , \mathbf{A}_i , and ν_i , respectively.

Herein, we consider parsimonious families of multivariate skew-normal and skew- t distributions that are non-elliptical, robust extensions to the GPCM family; we refer to our families as the SNCLUST and *St*CLUST families, respectively. Note that the geometric interpretation of the component shapes for members of the SNCLUST and *St*CLUST families is not the same as for members of the GPCM, MCLUST, and *t*EIGEN families unless the skewness is zero. In the GPCM family, the constraints (cf. Table 1) are applied to the decomposed component covariance matrices and so can be considered in terms of the volume, shape, and orientation of the component densities. However, for the SNCLUST and *St*CLUST families, we are decomposing the component scale matrices Ω_i ; accordingly, the shapes of the corresponding component densities cannot be interpreted after the fashion of MCLUST (or GPCM) unless the skewness is zero (cf. Wang et al., 2009). This lack of geometric interpretability is of no importance because the rationale for imposing constraints on the decomposed component scale matrices herein is simply to introduce parsimony.

3.2 Parameter Estimation

This section outlines the parameter estimation for the *St*CLUST family. Estimates for the *SN*CLUST family can be derived in a similar fashion. First, note that, from the representation given in (2), the joint distribution of \mathbf{y}, u, w can be written

$$\begin{aligned} f(\mathbf{y}, u, w) &= f(\mathbf{y}|u, w)f(u|w)f(w) \\ &= \frac{|w|^{1/2}}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\xi} - \boldsymbol{\lambda}|u|)' \left(\frac{\boldsymbol{\Omega}}{w} \right)^{-1} (\mathbf{y} - \boldsymbol{\xi} - \boldsymbol{\lambda}|u|) \right\} \\ &\quad \times \frac{1}{\sqrt{2\pi(1/w)}} \exp \left\{ -\frac{u^2}{2(1/w)} \right\} \frac{(\frac{\nu}{2})^{\nu/2}}{\Gamma(\frac{\nu}{2})} w^{\frac{\nu}{2}-1} \exp \left\{ -\frac{\nu}{2}w \right\} \\ &= \frac{(\frac{\nu}{2})^{\nu/2} w^{(\nu+p-1)/2}}{(2\pi)^{\frac{p+1}{2}} |\boldsymbol{\Omega}|^{\frac{1}{2}} \Gamma(\frac{\nu}{2})} \exp \left\{ -\frac{w}{2} [(\mathbf{y} - \boldsymbol{\xi} - \boldsymbol{\lambda}|u|)' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\xi} - \boldsymbol{\lambda}|u|) + u^2 + \nu] \right\}. \end{aligned}$$

Within the EM framework, it is convenient to adopt the following notation. We regard the observed data $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_n)$ as being incomplete, where $\mathbf{u} = (u_1, \dots, u_n)$ and $\mathbf{w} = (w_1, \dots, w_n)$ are unobservable latent variables and the component membership labels $\mathbf{z} = (z_1, \dots, z_n)$ are the missing data. Note that \mathbf{z} is defined such that $\mathbf{z}_j = (z_{1j}, \dots, z_{gj})'$ for $j = 1, \dots, n$, where $z_{ij} = 1$ if \mathbf{y}_j belongs to component i and $z_{ij} = 0$ otherwise. Now, $(\mathbf{y}, \mathbf{u}, \mathbf{w}, \mathbf{z})$ is referred to as the complete-data and the complete-data log-likelihood for unconstrained component scale matrices $\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_g$ is given by

$$\log \mathcal{L}(\boldsymbol{\Psi} | \mathbf{y}, \mathbf{u}, \mathbf{w}, \mathbf{z}) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log \pi_i + \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log f_i(\mathbf{y}_j, u_j, w_j) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3, \quad (6)$$

say, where

$$\begin{aligned} \mathcal{L}_1 &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log \pi_i, \\ \mathcal{L}_2 &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \left\{ -\frac{1}{2} \left[\log(2\pi) + \log |\boldsymbol{\Omega}_i^{-1}| + w_j (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j)' \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j) \right] \right\}, \\ \mathcal{L}_3 &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \left\{ -\frac{1}{2} [(p-1) \log(w_j) + w_i u_j^2] + \frac{\nu_i}{2} [w_j - \log(\nu_i/2)] + \log \Gamma(\nu_i/2) + (\nu_i/2 - 1) \log(w_j) \right\}. \end{aligned}$$

In the E-step, the expected value of the complete-data log-likelihood can be found using any of the three methods mentioned in Section 2; we use the method of Vrbik and McNicholas (2012) in our analyses (Sections 4 and 5). The updates for π_i , $\boldsymbol{\xi}_i$, $\boldsymbol{\lambda}_i$, and ν_i needed in the M-step are given by Wang et al. (2009); parameter estimation for the decomposed elements of $\boldsymbol{\Omega}_i = \boldsymbol{\lambda}_i \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i'$ will depend on the model under consideration (cf. Table 1) and can be found in an analogous fashion to the Gaussian case. To acquire the updates for λ_i , \mathbf{D}_i , and \mathbf{A}_i , we need only maximize \mathcal{L}_2 in (6), which is equivalent to minimizing

$$\begin{aligned} &\sum_{i=1}^g \sum_{j=1}^n z_{ij} \log |\boldsymbol{\Omega}_i^{-1}| + \sum_{i=1}^g \sum_{j=1}^n z_{ij} w_j (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j)' \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j) \\ &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log |\boldsymbol{\Omega}_i^{-1}| + \text{tr} \left\{ \sum_{i=1}^g \sum_{j=1}^n z_{ij} w_j (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j)' \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j) \right\} \\ &= \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log |\boldsymbol{\Omega}_i^{-1}| + \text{tr} \left\{ \sum_{i=1}^g W_i \boldsymbol{\Omega}_i^{-1} \right\}, \end{aligned}$$

where $W_i = \sum_{j=1}^n z_{ij} w_j (\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j)(\mathbf{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\lambda}_i u_j)'$. Finally, define $W := \sum_{i=1}^g W_i$ and note that parameter estimation for λ_i , \mathbf{D}_i , and \mathbf{A}_i becomes identical to that carried out by Celeux and Govaert (1995).

3.3 Model Selection

The best model from amongst the ten members of the *St*CLUST family is determined by the popular Bayesian information criterion (BIC; Schwarz, 1978). Although the underlying regularity conditions for the asymptotic approximation are not generally satisfied (cf. Keribin, 2000), the BIC has shown to be a useful tool for selecting amongst mixture models (e.g., Fraley and Raftery, 2002; Andrews and McNicholas, 2011a). The BIC is given by $-2l(\mathbf{x}, \hat{\boldsymbol{\vartheta}}) + m \log n$, where m is the number of free parameters, $l(\mathbf{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, and $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$. Dasgupta and Raftery (1998) propose using the BIC for mixture model selection. When defined as above, the model with the largest BIC is selected. Model selection for the SNCLUST family is handled in the same way.

3.4 Performance Assessment

Because the true classes for the data sets used in the applications that follow are known, the adjusted Rand index (ARI; Hubert and Arabie, 1985) is used to assess the clustering results. The ARI is a popular measure of classification agreement between the true and predicted group memberships or, more generally, between any two partitions. The ARI is a corrected form of the Rand index (Rand, 1971) that adjusts for chance agreement. An ARI of 1 corresponds to perfect agreement whereas an ARI of 0 corresponds to results no better than would be expected by guessing.

4 Simulation Studies

4.1 Introduction

Although the proposed methods can explicitly account for skewed components, they should also be capable of capturing symmetric components, e.g., data from a multivariate Gaussian or *t*-distribution. In Section 4.2, we illustrate that the *St*CLUST and SNCLUST families can uncover underlying Gaussian and *t*-mixtures. We then demonstrate the difficulties encountered by the MCLUST and *t*EIGEN families when faced with skewed data (Section 4.3). The `mclust` and `teigen` packages are used to run the MCLUST and *t*EIGEN models, respectively. In Section 4.4, we present a more difficult data set in which we compare the four algorithms. Note that to facilitate direct comparison with MCLUST, we restrict the *t*EIGEN, SNCLUST, and *St*CLUST families to ten models to correspond to the MCLUST covariance structures (cf. Table 1) for all our simulations.

4.2 Simulation 1

For this simulation, the first component, of size $n_1 = 300$, was generated from a bivariate normal distribution with parameters

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and the second component, of size $n_2 = 200$, was generated from a bivariate *t*-distribution with $\nu = 4$ degrees of freedom and parameters

$$\boldsymbol{\mu}_2 = \begin{pmatrix} -2 \\ 4 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The clustering performance of the *St*CLUST family was excellent, yielding an average ARI of 0.99 with standard deviation 0.007. The average fitted values for degrees of freedom were $\hat{\nu}_1 = 56.92$ and $\hat{\nu}_2 = 6.51$

and, looking across all runs, sensible values were almost always selected (Figure 2). The average values for skewness were $\hat{\lambda}_1 = (0.021, -0.035)$ and $\hat{\lambda}_2 = (0.014, 0.011)$, which makes sense when fitting elliptical components.

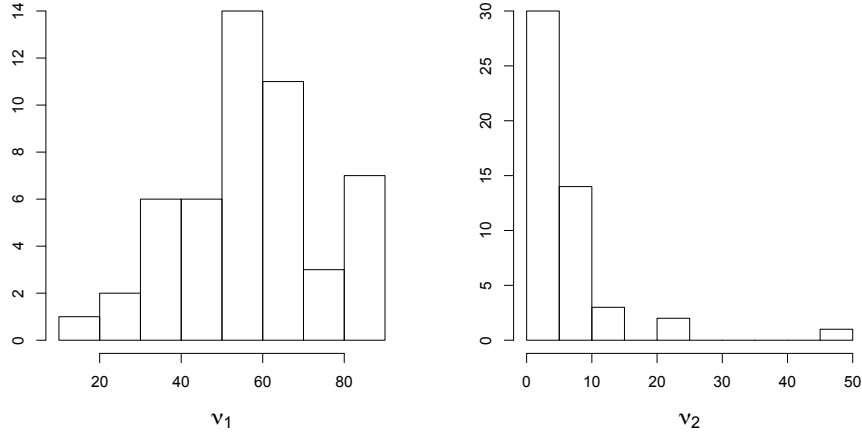


Figure 2: Histogram of the estimates for the degrees of freedom parameter from fitting the *St*CLUST models on the 50 random runs for Simulation 1.

The clustering performance of the SNCLUST family was also very good despite the fact that one of the components was simulated from a multivariate *t*-distribution.

4.3 Simulation 2

In this section, skewed data were generated from bivariate skew-*t* distributions using the convenient representation given in (2). In all, we used 50 random simulations with three components of sizes $n_1 = 150$, $n_2 = 200$, and $n_3 = 150$, respectively, with locations $\xi_1 = (0, 0)$, $\xi_2 = (6, 25)$, $\xi_3 = (4, 0)$, skewness vectors $\lambda_1 = (2, 4)$, $\lambda_2 = (-2, 4)$, $\lambda_3 = (2, 4)$, degrees of freedom $\nu_1 = 10$, $\nu_2 = 8$, $\nu_3 = 70$, and scale matrices

$$\Omega_1 = \begin{pmatrix} 0.4 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}, \quad \Omega_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \Omega_3 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.3 \end{pmatrix}.$$

Note that the third group has high degrees of freedom and is effectively a skew-normal cluster. All families were run for $i = 1, \dots, 9$ components; as a matter of practice, when the BIC chose a $g = 9$ component model, the algorithm was rerun for $i = 1, \dots, 12$.

When applied to these simulated data sets, the *St*CLUST family of models gave excellent clustering performance, obtaining an average ARI of 0.983 and correctly identifying the number of components 90% of the time (Table 2). MCLUST and *t*EIGEN, on the other hand, consistently overestimated the number of components and produced inferior ARI scores. The poor classification performance of the MCLUST and *t*EIGEN mixtures can sometimes be mitigated by merging components. For example, if we look at the results from a single run in Figure 3, we see that the classification performance of *t*EIGEN can be made equal to that of *St*CLUST by merging components. However, merging the MCLUST components here will not make its classification performance equal to that of *St*CLUST. The issue of merging is further considered in Sections 5.3 and 6.4.

Table 2: The number of components selected by MCLUST, *t*EIGEN, and *St*CLUST for simulation 2, along with average ARI values and standard deviations.

Groups Selected	MCLUST	<i>t</i> EIGEN	<i>St</i> CLUST
3		18%	90%
4	18%	32%	6%
5	18%	18%	4%
6	26%	20%	
7	30%	12%	
8	2%		
9	4%		
10	2%		
Mean ARI	0.747	0.749	0.983
Std. Dev. ARI	0.116	0.119	0.040

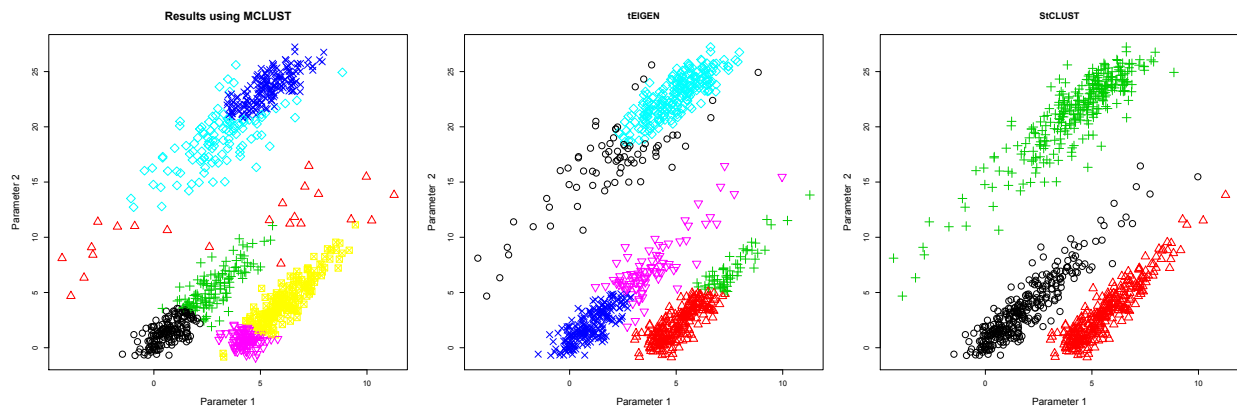


Figure 3: The maximum *a posteriori* classification obtained from fitted MCLUST, *t*EIGEN, and *St*CLUST models on one run of simulation 2.

4.4 Simulation 3

Both simulations in the previous sections represented well separated groups in two dimensions. To test these methods for a more complicated data set, a final simulation study was conducted using two overlapping groups in three dimensions. A two component multivariate skew-*t* data set was simulated with 100 observations in each group. Typically, the overlapping groups created an X-shaped figure (e.g., Figure 4). One-hundred such random data sets were simulated and the results (Table 3) show that the SNCLUST and *St*CLUST models perform at least as well as the MCLUST and *t*EIGEN families. Notably, *St*CLUST selected a two-component model for 98% of the simulated data sets and obtained the highest average ARI. Figure 5 plots the clustering results corresponding to simulated data in Figure 4.

5 Applications

5.1 Data Sets

The efficacy of the *St*CLUST and SNCLUST models is demonstrated on four real data sets. The **bank**, **crabs**, **iris**, and **wine** data are frequently used benchmark data sets for testing the performance of various model-

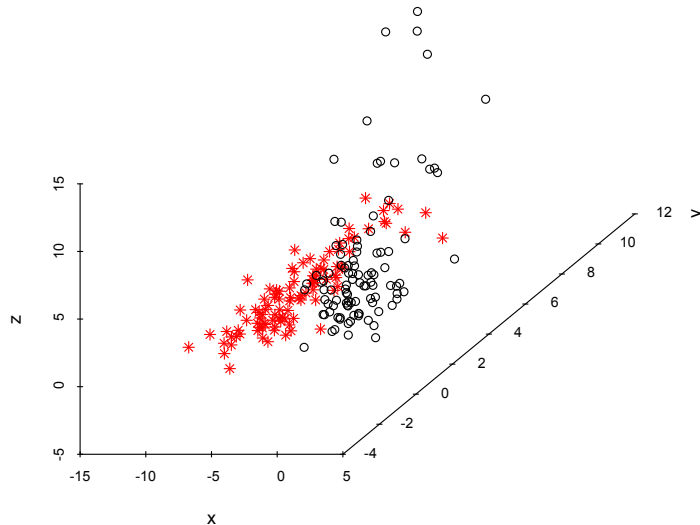


Figure 4: A typical 3D scatterplot of the simulation as described in Section 4.4, where \circ represents an observation from simulated component 1 and $*$ represents an observation from a simulated component 2.

Table 3: The number of components selected by MCLUST, t EIGEN, and St CLUST for simulation 3, along with average ARI values and standard deviations.

Groups Selected	MCLUST	t EIGEN	SNCLUST	St CLUST
2	6%	64%	60%	98%
3	51%	12%	38%	2%
4	35%	16%	2%	
5	8%	7%		
6		1%		
Mean ARI	0.482	0.572	0.606	0.650
Std. Dev. ARI	0.106	0.134	0.090	0.073

based clustering algorithms. As in Section 4, we facilitate direct comparison with MCLUST by restricting the t EIGEN, SNCLUST, and St CLUST families to analogues of the ten MCLUST models for all analyses in Section 5.

The Swiss Bank data (Flury and Riedwyl, 1988) records six different measurements on 100 genuine and 100 counterfeit Swiss banknotes. The crabs data (Campbell and Mahon, 1974) contain the morphological measurements of frontal lobe size, rear width, carapace length, carapace width, and body depth. The measurements are taken in millimetres on 200 crabs, 50 from each combination of sex (male, female) and colour (blue, orange). The iris data (Anderson, 1935; Fisher, 1936) contain the length and width in centimetres of the sepal and petal of three species of irises. The data set comprises the measurements of 150 flowers, 50 from each of the various species: *setosa*, *versicolor*, and *virginica*. The wine data (Forina et al., 1988) contain 13 chemical and physical properties of wine. The data set comprises measurements from 178 wines from three different cultivars: Barolo, Barbera, and Grignolino.

5.2 Results

The MCLUST, t EIGEN, SNCLUST, and St CLUST families — with the latter three restricted to the MCLUST models — were fitted to the scaled data for $i = 1, \dots, 9$. Again, in the event that a $g = 9$

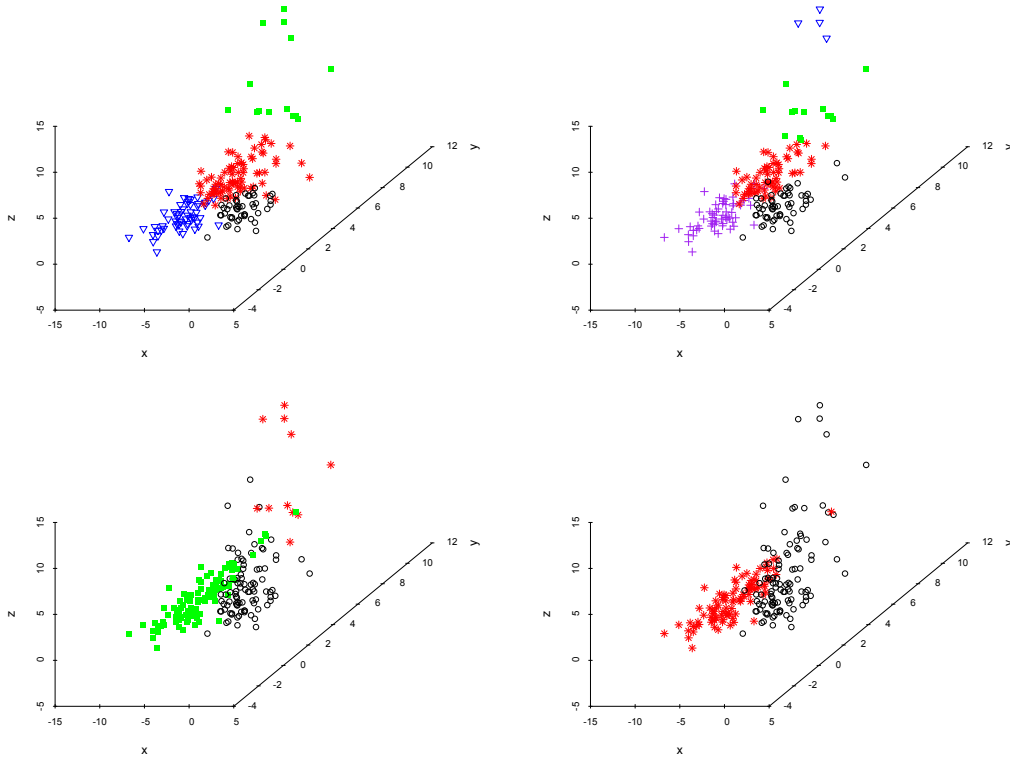


Figure 5: The solutions found by MCLUST, t EIGEN, SNCLUST, and St CLUST (top left, top right, bottom left, bottom right, respectively) when applied to the simulated data presented in Figure 4.

component model was selected, models were rerun for $i = 1, \dots, 12$. The chosen models (in accordance with the BIC) are presented with their corresponding ARIs in Table 4; at least one of our two skewed families outperforms both MCLUST and t EIGEN in each case.

Table 4: The ARI and members selected for each family when run on the bank, crabs, iris, and wine data sets. For each data set, the best ARI is highlighted in bold face.

	Model Selected, Number of Components, ARI			
	MCLUST	t EIGEN	SNCLUST	St CLUST
Bank	EEE 4, 0.679	CCCC, 4, 0.679	EEE, 4, 0.606	EEE, 2, 0.980
Crabs	EEE, 4, 0.311	CCCC, 10, 0.508	EEE, 4, 0.838	EEE, 5, 0.741
Iris	VVV, 2, 0.568	UUUC, 2, 0.568	VEI, 3, 0.941	VEI, 3, 0.922
Wine	VEI, 8, 0.481	CICC, 7, 0.454	VEI, 4, 0.801	EVI, 3, 0.947

Focusing first on the **bank** data set, we see comparable results between the MCLUST, t EIGEN, and SNCLUST families. However, a great improvement in ARI is achieved when the St CLUST family is used. The superiority of St CLUST over the other three families is largely attributable to the fact that it selects the same number of components as true classes. The remaining families select a four-component model and so their performance might be improved via merging components (Section 5.3).

The famous **crabs** data set is typically classified by sex and species (blue or orange), corresponding to four groups. Although the MCLUST family correctly selects a four-component model, the group memberships do not agree with this known partition and produced a low ARI (0.311). The t EIGEN family obtained a

higher ARI (0.508); however, it greatly overestimates the number of groups. SNCLUST, which selects a four-component model, performs much better than its competitors. This is not surprising because StCLUST is indicating skew-normal components ($\hat{\nu}_1 = 90, \hat{\nu}_2 = 89.807, \hat{\nu}_3 = 90, \hat{\nu}_4 = 8.455, \hat{\nu}_5 = 90, \hat{\nu}_6 = 90$) and the parameter estimates suggest that the groups are negatively skewed; $\hat{\lambda}_1 = (-1.4, -1.2, -1.5, -1.6, -1.4)'$, $\hat{\lambda}_2 = (-1.9, -1.5, -2.0, -2.0, -1.9)'$, $\hat{\lambda}_3 = (-1.3, -1.7, -1.4, -1.5, -1.3)'$, $\hat{\lambda}_4 = (-1.3, -1.3, -1.2, -1.2, -1.2)'$.

For the *iris* data, the ARI obtained for both the MCLUST models and their *t*-analogues are comparable. Thus, little gain is achieved through the inclusion of the degrees of freedom parameter present in the *t*EIGEN and StCLUST models. The failure of the addition of the degrees of freedom parameter to improve on the clustering results here might be connected to its limitations as a one-dimensional parameter. Had we used a *p*-dimensional degrees of freedom parameter, allowing different degrees of freedom in each dimension, some improvement may have been observed. This extension will be explored in future work.

The results for the *wine* data set tell a similar story to that of the banknotes. As can be seen in Table 4, MCLUST and *t*EIGEN produce almost identical clustering results; however, a considerable improvement in ARI is apparent once skewness is introduced. As mentioned previously, the increased ARI is greatly owing to the fact that StCLUST selects the value of *g* corresponding to the number of true classes.

5.3 Comparing with Merged Clusters

The recent work of Baudry et al. (2010) argues that the number of mixture components does not necessarily correspond to the number of true groups or clusters. For instances where an underlying group is comprised of a mixture of two or more Gaussian distributions, the foregoing paper proposes a method for combining mixture components to represent one cluster. This method first fits a mixture of Gaussian distributions with *g* components then successively merges mixture components together, resulting in a *k* component solution where $k \leq g$. The recent version of `mclust` (i.e., version 4.0, Fraley et al., 2012) contains a function called `clustCombi` that implements the methodology proposed in Baudry et al. (2010). This section compares the clustering results produced by `clustCombi` with the results presented in Section 5.2.

To compare our results with the optimal solution derived by merging components, we used two different merging techniques. The first involves the `clustCombi` function, which starts with the original *g* component solution (as chosen by the BIC) and combines two components according to an entropy criterion to obtain a *g* - 1 component solution. This procedure is carried out until the one component solution is obtained. The ARI is calculated at each merging step and the solution with the largest ARI is saved as the ‘best’. Hereafter we refer this result as the maximum `clustcombi` solution or MCC for short. The second method involves merging components by hand to create the most advantageous solution. When looking at a cross tabulation of the known group labels (or true clusters) against the clustering results obtained from the clustering algorithm, this becomes a matter of combining the columns that will maximize the agreements between group labels and components. An example of this merging process using cross-tabulations (or so-called classification tables) is demonstrated using the MCLUST solution for the *bank* dataset in Table 5. Herein, we shall refer to this second merging method as the merging by hand or MBH solution. Note that both the MCC and MBH procedures rely on knowing the true group membership labels; therefore, they could not be used for real clustering applications.

Table 6 summarizes the ARI values for the pertinent models (indicated using gray font) after performing the MCC and MBH methods described above. Note that if the true number of clusters is greater than or equal to the number of components found by the clustering algorithm, no gain in ARI can be obtained by merging. For instance, when the clustering algorithms under consideration are applied to the *iris* data set (which has three true clusters) either a two or three component solution is obtained and no combination of component merging will result in a better classification. Because no merging is done, the results are unchanged from Table 4 and are given in black font in Table 6. Even when a merging technique is used, the SNCLUST and StCLUST models perform as well as or better than the MCLUST and *t*EIGEN families in each case.

Table 5: Left: Classification results found by MCLUST cross-tabulated against the true group labels. Right: The merging by hand (MBH) classification results (created by merging the second with the third components found by MCLUST and similarly merging the first with the fourth) cross-tabulated against the true group labels.

True	MCLUST				merge MCLUST components 2 and 3 & components 1 and 4 →	True	MBH	
	1	2	3	4			1	2
1	1	75	24	0		1	1	99
2	15	0	0	85		2	100	0

Table 6: ARI values for each data set after merging components is considered. The ARI values calculated after merging components (cf. Table 4) are given in grey font. Where merging cannot improve classification performance, the (unchanged) ARI values are given in black font. For each ARI value, the associated number of components is given in brackets and for each data set, the best ARI value is typed in bold face.

	ARI				
	MCC	MBH			
	MCLUST	MCLUST	tEIGEN	SNCLUST	StCLUST
Bank	0.860 (3)	0.980 (2)	0.980 (2)	0.980 (2)	0.980 (2)
Crabs	0.311 (4)	0.679 (4)	0.783 (4)	0.838 (4)	0.753 (4)
Iris	0.568 (2)	0.568 (2)	0.568 (2)	0.941 (3)	0.922 (3)
Wine	0.876 (5)	0.901 (3)	0.867 (3)	0.947 (3)	0.947 (3)

6 Model-Based Classification

6.1 The Model

In the event that a subset of the data under consideration has known group membership labels, model-based classification can be used. This is a semi-supervised version of model-based clustering that uses both the labeled and unlabelled data to produce parameter estimates. If we order the data such that the first m observations ($m < n$) have known labels, then the Gaussian model-based classification likelihood is

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{z}_1, \dots, \mathbf{z}_m) = \prod_{j=1}^m \prod_{i=1}^g [\pi_i \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)]^{z_{ij}} \prod_{k=m+1}^n \sum_{l=1}^h \pi_l \phi(\mathbf{x}_k \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l),$$

for $h \geq g$; often, as in our analyses (Sections 6.2, 6.3, and 6.4), it is assumed that $h = g$.

We apply the SNCLUST and StCLUST families for model-based classification, drawing comparison with the MCLUST and tEIGEN families. In addition to the real data that we used to illustrate our clustering approaches, we analyze real food authenticity data on Italian olive oils. To facilitate direct comparison with MCLUST, we restrict the tEIGEN, SNCLUST, and StCLUST families to analogues of the ten models used in MCLUST (cf. Table 1) for the analyses in Sections 6.2 and 6.3. However, we consider all 14 models (i.e., all 14 constraints for the decomposition of the component scale matrices) for the SNCLUST and StCLUST families in Section 6.4.

6.2 Olive Oil Data

The olive oil data (Forina and Tiscornia, 1982; Forina et al., 1983) contain the percentage of eight fatty acids found in 572 Italian olive oils. The data set are available within the `pgmm` package (McNicholas et al., 2011) for R. Each oil sample comes from one of three distinct regions, which can be further partitioned into nine

areas. McNicholas (2010) used these data to illustrate model-based classification, by taking random subsets of 50% of the observations to have known component membership labels; we will do the same to illustrate our models.

We first consider the problem of classifying the olive oils into their regions. The four families were run using 15 different random subsets with 50% of the labels known and $g = 3$. As advocated by Andrews et al. (2011), we employ a uniform initialization whereby the unknown observations have initial predicted labels $\hat{z}_{ij} = 1/g$, for $i = 1, \dots, g$ and $j = 1, \dots, n$. A similar approach is repeated for the $g = 9$ component models, this time classifying the data by geographical area. The average ARI values for the fitted models (Table 7) confirm the slightly superior performance of the SNCLUST family. However, we note that all four families give good classification performance when classifying by region and by area.

We notice that the *St*CLUST models obtain a lower ARI than the *t*EIGEN models when classifying the olive oil into geographical area. One possible explanation is that uniform initialization is not suitable for the *St*CLUST family. Because these procedures are susceptible to converge to local maxima, and considering the relatively complicated likelihoods associated with the skew-*t* family, perhaps a more guileful approach such as a deterministic annealing EM (cf. Ueda and Nakano, 1998) is necessary. Further investigation of this idea will be a subject of future work.

Table 7: The average ARI values, with standard deviations, for the fitted MCLUST, *t*EIGEN, SNCLUST, and *St*CLUST models for the olive data by region ($g = 3$) and area ($g = 9$). For each data set, the best ARI is highlighted in bold face.

	MCLUST	<i>t</i> EIGEN	SNCLUST	<i>St</i> CLUST
Olive (Region)	0.9962 (0.0036)	0.9965 (0.0032)	0.9983 (0.0028)	0.9946 (0.0021)
Olive (Area)	0.9091 (0.0333)	0.9185 (0.0224)	0.9249 (0.0242)	0.8994 (0.0307)

6.3 Other Data

The same four data sets considered in Section 5 are used for model-based classification; again, 50% of the labels are taken to be known. Each data set was run with 25 different random subsets of known labels with the number of groups specified to the correct value. Although the MCLUST and *t*EIGEN families benefit substantially under a classification framework, SNCLUST obtained the highest ARI on all four data sets (Table 8). Comparing *St*CLUST with the its symmetric alternative *t*EIGEN, we see an improvement in ARI for all but the crabs dataset. As mentioned in Section 6.2, this inconsistency could be a result of the uniform initialization.

Table 8: Average ARI values, with standard deviations, for the fitted MCLUST, *t*EIGEN, SNCLUST, and *St*CLUST models. For each data set, the best ARI is highlighted in bold face.

	MCLUST	<i>t</i> EIGEN	SNCLUST	<i>St</i> CLUST
Bank	0.9760 (0.0200)	0.9760 (0.0200)	0.9911 (0.0102)	0.9792 (0.0204)
Crabs	0.8798 (0.0474)	0.8818 (0.0451)	0.9255 (0.0389)	0.8575 (0.0632)
Iris	0.8910 (0.0311)	0.8910 (0.0311)	0.9447 (0.0305)	0.9199 (0.0574)
Wine	0.9004 (0.0396)	0.8271 (0.0505)	0.9444 (0.0504)	0.8379 (0.0770)

6.4 The EVV, EVE, VEE and VVE models

Heretofore, we restricted the SNCLUST and *St*CLUST families to correspond to the ten MCLUST models. This restriction was appropriate to facilitate direct comparison with MCLUST. In this section, we will consider the full SNCLUST and *St*CLUST families; i.e., the SNCLUST and *St*CLUST families with all 14

models (Table 1). We apply the full SNCLUST and *St*CLUST families for model-based classification of the data sets considered in Sections 6.2 and 6.3. We compare the performance of the full families to the versions restricted to correspond to MCLUST. Note that we implement the four ‘extra’ models for the SNCLUST and *St*CLUST families using the majorization-minimization (MM) algorithm (cf. Hunter and Lange, 2004) developed by Browne and McNicholas (2013).

Table 9 presents ARI scores associated with the reduced (i.e., MCLUST-analogous) as well as the full SNCLUST and *St*CLUST families. The last two columns summarize the proportion of runs where one of the four ‘extra’ models was chosen (these correspond to the EVV, EVE, VEE, and VVE models in Table 1). Our results demonstrate that one of these models is often chosen, resulting in a higher average ARI score for all but one of the *St*CLUST runs as well as half of the SNCLUST runs. To be more specific, EVE was selected between 59% and 100% of the time for all but the olive oil classification by region, where the EVE model was never selected. We therefore advocate the inclusion of all fourteen models in practice rather than the subset of ten models used by MCLUST.

Table 9: Average ARI values, with standard deviations, for the fitted SNCLUST, and *St*CLUST using the ten model (original) and 14 model (full) families. For each data set, the best ARI is highlighted in bold face.

	Original 10 model family		Full 14 model family		% of non-MCLUST	
	SNCLUST	<i>St</i> CLUST	SNCLUST	<i>St</i> CLUST	SNCLUST	<i>St</i> CLUST
Bank	0.9911 (0.010)	0.9792 (0.020)	0.9908 (0.010)	0.9896 (0.010)	100%	100%
Crabs	0.9255 (0.039)	0.858 (0.063)	0.9109 (0.028)	0.9170 (0.033)	93%	92%
Iris	0.9447 (0.031)	0.9199 (0.057)	0.9622 (0.024)	0.9658 (0.020)	98%	100%
Wine	0.9444 (0.050)	0.8379 (0.077)	0.9374 (0.046)	0.9474 (0.030)	64%	85%
Olive(Region)	0.9981 (0.004)	0.9965(0.003)	0.9994 (0.002)	0.9983 (0.002)	100%	67%
Olive(Area)	0.9091 (0.033)	0.9185 (0.022)	0.9655 (0.012)	0.9562 (0.011)	100%	100%

7 Concluding Remarks

This paper builds on the growing trend towards non-Gaussian model-based clustering by developing two families of models that account for skewness: skew-normal and skew-*t* analogues of the GPCM family of models. Parameter estimation was outlined and our novel families were applied to simulated and real data. Both model-based clustering and classification were illustrated using several real data sets, and the performance of *St*CLUST and SNCLUST was generally superior to their symmetric analogues. Interestingly, this superior performance was often retained even after merging components was considered.

Future work will focus on the initialization of these families in addition to the search for more effective model selection techniques. As briefly mentioned in Section 5.2, we could investigate the efficacy of extending the degrees of freedom to be p -dimensional. From the classification viewpoint, we only considered the semi-supervised model-based classification scenario, where one mixture component corresponds to a class. The straightforward extension to model-based discriminant analysis will be investigated and incorporated into R packages that are being developed for the *St*CLUST and SNCLUST families. Finally, the use of our skewed families in the analysis of data of mixed type (cf. Browne and McNicholas, 2012) will also be investigated.

Acknowledgements

This work was supported by an Ontario Graduate Scholarship, a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and an Early Researcher Award from the Ontario Ministry of Research and Innovation.

References

- Anderson, E. (1935). The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society* 59, 2–5.
- Andrews, J. L. and P. D. McNicholas (2011a). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing* 21(3), 361–373.
- Andrews, J. L. and P. D. McNicholas (2011b). Mixtures of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis. *Journal of Statistical Planning and Inference* 141(4), 1479–1486.
- Andrews, J. L. and P. D. McNicholas (2012a). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing* 22(5), 1021–1029.
- Andrews, J. L. and P. D. McNicholas (2012b). *teigen: Model-based clustering and classification with the multivariate t-distribution*. R package version 1.0.
- Andrews, J. L., P. D. McNicholas, and S. Subedi (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis* 55, 520–529.
- Baek, J. and G. J. McLachlan (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualisation of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 1298–1309.
- Baek, J. and G. J. McLachlan (2011). Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics* 27(9), 1269–1276.
- Banfield, J. D. and A. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49, 803–821.
- Baudry, J., A. Raftery, G. Celeux, K. Lo, and R. Gottardo (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* 19(2), 332–353.
- Bensmail, H. and G. Celeux (1996). Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association* 91, 1743–1748.
- Bouveyron, C., S. Girard, and C. Schmid (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis* 52(1), 502–519.
- Browne, R. P. and P. D. McNicholas (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference* 142(11), 2976–2984.
- Browne, R. P. and P. D. McNicholas (2013). Estimating common principal components in high dimensions. Arxiv preprint.
- Cabral, C., V. Lachos, and M. Prates (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis* 56(1), 126–142.
- Campbell, N. A. and R. J. Mahon (1974). A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian Journal of Zoology* 22, 417–425.
- Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition* 28, 781–793.
- Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93, 294–302.

- Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society: Series C* 55(1), 1–14.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1), 1–38.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(Part II), 179–188.
- Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.
- Forina, M., C. Armanino, S. Lanteri, and E. Tiscornia (1983). Classification of olive oils from their fatty acid composition. *Food Research and Data Analysis*, 189–214.
- Forina, M., R. Leardi, C. Armanino, S. Lanteri, and B. Vandeginste (1988). Parvus: An extendable package of programs for data exploration, classification and correlation. *Journal of Chemometrics* 4(2), 191–193.
- Forina, M. and E. Tiscornia (1982). Pattern recognition methods in the prediction of italian olive oil origin by their fatty acid content. *Annali di Chimica* 72, 143–155.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.
- Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington.
- Frühwirth-Schnatter, S. and S. Pyne (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* 11(2), 317–336.
- Ho, H., S. Pyne, and T. Lin (2012). Maximum likelihood inference for mixtures of skew student-t-normal distributions through practical EM-type algorithms. *Statistics and Computing* 22(1), 287–299.
- Ho, H. J., T.-I. Lin, H. Chen, and W. L. Wang (2012). Some results on the truncated multivariate t distribution. *Journal of Statistical Planning and Inference* 142, 25–40.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
- Hunter, D. L. and K. Lange (2004). A tutorial on MM algorithms. *The American Statistician* 58(1), 30–37.
- Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing* 19(1), 73–83.
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A* 62(1), 49–66.
- Kotz, S. and S. Nadarajah (2004). *Multivariate t distributions and their applications*. Cambridge University Press.
- Lee, S. and G. McLachlan (2012). Finite mixtures of multivariate skew t -distributions: Some recent and new results. *Statistics and Computing*. To appear.
- Lin, T. I., J. C. Lee, and S. Y. Yen (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica* 17, 909–927.

- McLachlan, G., R. Bean, and L. Ben-Tovim Jones (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t distribution. *Computational Statistics and Data Analysis* 51, 5327–5338.
- McLachlan, G. J. and K. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- McLachlan, G. J. and D. Peel (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, Volume 1451, pp. 658–666. Berlin: Springer-Verlag.
- McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference* 140(5), 1175–1181.
- McNicholas, P. D., K. R. Jampani, A. F. McDaid, T. B. Murphy, and L. Banks (2011). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.0.
- McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing* 18(3), 285–296.
- McNicholas, P. D. and T. B. Murphy (2010a). Model-based clustering of longitudinal data. *Canadian Journal of Statistics* 38(1), 153–168.
- McNicholas, P. D. and T. B. Murphy (2010b). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* 26(21), 2705–2712.
- McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference* 142(5), 1114–1127.
- Peel, D. and G. J. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* 10(4), 339–348.
- Pyne, S., X. Hu, K. Wang, E. Rossin, T. I. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. L. De Jager, and J. P. Mesirov (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences USA* 106, 8519–8524.
- R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.
- Sahu, S. K., D. Dey, and M. Branco (2003). A new class of multivariate skew distributions with application to Bayesian regression models. *Canadian Journal of Statistics* 31, 129–150.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.
- Steane, M. A., P. D. McNicholas, and R. Yada (2012). Model-based classification via mixtures of multivariate t-factor analyzers. *Communications in Statistics – Simulation and Computation* 41(4), 510–523.
- Ueda, N. and R. Nakano (1998). Deterministic annealing EM algorithm. *Neural Networks* 11, 271–282.
- Vrbik, I. and P. D. McNicholas (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics & Probability Letters* 82(6), 1169–1174.
- Wang, K., S. Ng, and G. McLachlan (2009). Multivariate skew t mixture models: Applications to fluorescence-activated cell sorting data. In H. Shi, Y. Zhang, M. Bottema, B. Lovell, and A. Maede (Eds.), *Digital Image Computing: Techniques and Applications, 2009. DICTA '09.*, pp. 526–531. IEEE.